On Statistical Reproducibility of Normality and Equality of Variances Tests

Norah D. Alshahrani[1†], Tahani Coolen-Maturi[2*†], Frank P. A. Coolen[2†]

[1]Department of Mathematics, University of Bisha, Bisha, 67721, Saudi Arabia.
[2*]Department of Mathematical Sciences, Durham University, Durham, DH1 3LE, UK.

*Corresponding author(s). E-mail(s): tahani.maturi@durham.ac.uk;
Contributing authors: ndmuflih@ub.edu.sa; frank.coolen@durham.ac.uk;
†These authors contributed equally to this work.

Abstract

Reproducibility of research is crucial and has received much attention in recent years. One aspect of reproducibility is statistical reproducibility, which examines whether statistical inferences remain similar when experiments are repeated. This paper investigates the reproducibility probability (RP) of normality tests and tests for equality of variances from a nonparametric predictive inference (NPI) perspective, offering a novel predictive framework for quantifying test reproducibility without strong parametric assumptions. Three well-known normality tests—the Shapiro-Wilk (SW), Anderson-Darling (AD), and Lilliefors (LF) tests—are studied, along with two tests for equality of variances: the F-test and Levene's test. The results show that RP tends to be low, particularly when p-values are close to the significance threshold. RP is also influenced by sample size and significance level, with larger samples decreasing RP in the non-rejection area and increasing it in the rejection area. Among the normality tests, the Shapiro-Wilk test has the highest RP in the non-rejection area, while the Anderson-Darling test has the highest RP in the rejection area. For the equality of variances tests, the F-test exhibits greater variability, particularly under non-normality. These findings highlight the limited statistical reproducibility of widely used tests and demonstrate how the proposed NPI-based approach can provide practical insight into the stability of test outcomes under uncertainty.

Keywords: Reproducibility probability (RP), Nonparametric predictive inference (NPI), Normality tests, Equality of variances tests

1 Introduction

The reproducibility of statistical hypothesis tests has received increasing attention in recent years. A fundamental principle of scientific credibility is the ability to obtain similar results when a hypothesis test is repeated under the same sample size and conditions. Reproducible results confirm that inferences made from statistical tests are reliable and not merely the outcome of chance or specific analytical choices. Despite its crucial role in science, reproducibility lacks a universally accepted definition, and its distinction from related concepts such as replicability remains ambiguous. The literature presents multiple, sometimes conflicting, definitions, leading to inconsistencies in discussions on statistical reliability. According to the National Academies of Sciences, Engineering, and Medicine [30], *reproducibility* refers to the ability to obtain consistent results using the same data and analytical methods, whereas *replicability* refers to the ability to achieve consistent findings in a new, independent study using different data but similar methods.

Goodman [23] was among the first to highlight concerns about the reproducibility of statistical findings, emphasising that p-values are often misinterpreted, giving a misleading impression of confidence in research conclusions. He argued that p-values do not measure effect size or reproducibility probability and called for a more transparent approach to statistical inference that incorporates additional metrics. Senn [32] expanded on these ideas, distinguishing between reproducibility probability and the p-value. While Senn questioned Goodman's claim that p-values overstate evidence against the null hypothesis, he acknowledged a relationship between p-values and reproducibility.

Beyond statistical inference, the broader concept of reproducibility has been the focus of significant discussion in the scientific community, particularly regarding its challenges and best practices. Atmanspacher and Maasen [4] provided an overview of key issues such as publication bias, the importance of following best practices, and the challenges researchers face in improving reproducibility. However, surprisingly little attention has been paid to the reproducibility of statistical inference methods themselves, despite their critical role in empirical research [15].

The work of Simkus et al. [36] provides a detailed examination of the ambiguity surrounding reproducibility, classifying existing definitions into five distinct types and analysing the impact of variations in datasets, laboratories, and experimental conditions that influence reproducibility. The paper also examines statistical reproducibility, noting that, like reproducibility itself, it lacks a clear definition. Goodman [23] described it as the probability of obtaining another statistically significant result in the same direction under identical conditions, but this perspective is just one among many. Simkus et al. [36] review statistical approaches to quantifying reproducibility, the role of p-values, and the challenges posed by variability across studies. A key contribution is the argument that reproducibility should be treated as a predictive problem, and Nonparametric Predictive Inference (NPI) is proposed as a framework for addressing this issue. Finally, the paper highlights ethical concerns in preclinical research and calls for further work on decision-making strategies when reproducibility is low.

Building on these discussions, this paper investigates the reproducibility of statistical hypothesis tests through a predictive lens. Prior research on statistical reproducibility has explored methods such as power-based estimation of reproducibility probability , and distinctions between different types of test repetition [19–22]. Miller [29] emphasised the importance of distinguishing between two types of repetition: (1) repetition by independent researchers under different conditions and (2) repetition by the same researcher under identical conditions. While his skepticism about drawing conclusions from independent replications when the true effect size and test power were unknown is noteworthy, this paper focuses on the second scenario. Here, statistical reproducibility refers to the probability that the same test outcome would be reached if the test were repeated under the same conditions.

This study employs NPI, a frequentist approach with minimal assumptions, to assess test reproducibility. NPI has been successfully applied to various statistical problems, including diagnostic accuracy analysis [18], finance [6], and operations research [14]. Its predictive nature makes it well-suited for reproducibility studies, where the goal is to assess the likelihood of obtaining the same test outcome in future experiments. Coolen and BinHimd [12, 13] pioneered the application of NPI in reproducibility studies, investigating its use for nonparametric tests such as the Wilcoxon Mann–Whitney test, Signed-Rank test, and the Kolmogorov–Smirnov test. Subsequent work extended NPI reproducibility assessment to population quantiles [10, 11] and t-tests in pharmaceutical applied settings [35].

In this paper, we examine the reproducibility of two fundamental classes of statistical tests: normality tests, which assess whether a dataset follows a normal distribution, and equality of variances tests, which evaluate whether two groups have equal variances. These tests are crucial for the validity of many parametric analyses, and understanding their reproducibility under different conditions—such as varying sample sizes and distributions—is essential.

While it is well-known that many commonly used tests for normality and equality of variances exhibit sensitivity to deviations from assumptions, the key contribution of this study is to systematically evaluate their reproducibility using the nonparametric predictive inference (NPI) framework. To our knowledge, this is the first application of NPI-based reproducibility probability (NPI-RP) to these pre-tests. Although these tests are widely employed in practice, their reproducibility—an important dimension of scientific reliability—has not been thoroughly examined in this context. Our analysis provides a novel perspective on the stability of statistical decisions under repeated sampling, offering valuable insights for applied researchers.

The remainder of this paper is structured as follows. Section 2 provides an overview of the predictive approach to statistical reproducibility, first introduced by Coolen and BinHimd [12], within the NPI framework. It discusses the benefits of viewing reproducibility as a predictive problem and highlights challenges associated with traditional reproducibility assessments. Section 3 investigates the reproducibility of normality tests, while Section 4 explores the reproducibility probability of equality of variances tests. Finally, Section 5 summarises the study and discusses potential directions for future research.

2 Statistical Reproducibility

2.1 A Predictive Framework for Statistical Reproducibility

A novel perspective on reproducibility probability (RP) was introduced by Coolen and BinHimd [12] through the application of Nonparametric Predictive Inference (NPI), a framework rooted in frequentist statistical methods. The predictive nature of NPI naturally lends itself to assessing RP, as it enables the prediction of future test outcomes based on an initial test, assuming identical conditions and sample size [8]. Coolen and BinHimd [12, 13] pioneered the use of NPI for reproducibility assessment, applying it to basic nonparametric tests such as the Wilcoxon–Mann–Whitney test (WMT), the Signed-Rank test, and the Kolmogorov–Smirnov test. Alqifari and Coolen [1, 11] extended this approach to tests involving population quantiles and precedence tests. Additionally, Simkus et al. [35] explored NPI reproducibility in the context of multiple t-tests, emphasising the challenges posed by multiple testing.

In the Bayesian framework, Billheimer [7] advocated for predictive inference by focusing on the prediction of future observable data rather than inferring unobservable parameters to enhance reproducibility. While this approach aligns with the goal of statistical reproducibility, this paper employs the NPI framework as a preferable alternative, as it avoids distributional assumptions about the data.

The NPI-based method for assessing reproducibility involves conducting a test on the original data and evaluating its results across all possible future data sets of the same size, assuming post-data exchangeability [15]. Reproducibility is then assessed by deriving lower and upper bounds for the reproducibility probabilities, utilising Hill's assumption, which will be discussed in more detail below.

Senn [32] argued that, in the worst-case scenario, the probability of reproducibility of a hypothesis test could be as low as 0.5, particularly when the test statistic is near the threshold between rejecting and failing to reject the null hypothesis. Coolen and BinHimd [12] confirmed this result for basic tests involving a single group of data, where the minimum NPI lower reproducibility probability was found to be 0.5 [15]. However, when testing with two groups of data, the minimum lower reproducibility probability was found to be lower than 0.5, and reproducibility tended to decrease further if the null hypothesis was rejected with a test statistic close to the rejection threshold. This issue is further exacerbated by the nature of hypothesis testing, which is often designed to maximise the likelihood of rejecting the null hypothesis—an objective commonly found in experimental studies [15]. Additionally, concerns have been raised that both lower and upper NPI reproducibility probabilities may remain low even for test statistics that are far from their respective thresholds [15].

Although this approach provides valuable insights, it can become computationally intensive for complex tests or large datasets [15]. When it is possible to determine whether a given ordering of future observations among the original data will lead to rejection or non-rejection of the null hypothesis—without assuming specific values between two original observations—sampling future orderings provides a computationally feasible solution. This method enables the estimation of both lower and upper reproducibility probabilities within the NPI framework [15, 16]. In cases where precise knowledge of future observations is required to determine the outcome, the NPI bootstrap method [8, 13, 15] is recommended.

This paper employs the nonparametric predictive inference bootstrap (NPI-B) to estimate reproducibility probability (NPI-B-RP). While NPI-B provides a point estimate of reproducibility probability, it does not offer an explicit means of expressing results in terms of imprecise reproducibility probabilities. NPI-B-RP leverages the NPI-B framework for prediction, estimating RP by repeatedly conducting a statistical test and examining the consistency of test outcomes [8]. BinHimd [8] performed a preliminary study on NPI-B-RP for the Wilcoxon Mann–Whitney test, demonstrating that NPI-B-RP yields results consistent with theoretical values for both lower and upper reproducibility probabilities, even with small sample sizes. Similarly, Simkus [34] applied the NPI-B-RP algorithm to the t-test, modifying certain aspects, such as the number of bootstrap iterations and the reporting of NPI-B-RP using different summary statistics.

In this paper, we use the NPI Bootstrap (NPI-B) technique to assess reproducibility for normality tests and equality of variance tests. NPI-B applies bootstrap resampling to estimate reproducibility probability, making it particularly useful when exact analytical solutions are impractical. By simulating repeated samples from the observed data, NPI-B provides a practical approach to evaluating reproducibility across different statistical tests.

In the next section, we will explore the mathematical foundations of NPI and NPI-B in more detail, demonstrating their application to the problem of reproducibility in statistical experiments.

2.2 Nonparametric Predictive Inference Bootstrap (NPI-B)

Nonparametric predictive inference (NPI) [5, 14] is a frequentist statistical method based on Hill's assumption $A_{(n)}$ [24]. This assumption allows for direct probabilistic predictions of future observations, conditional on the observed data. Suppose we have n ordered real-valued observations, $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$, corresponding to continuous and exchangeable random variables $X_1, X_2, \ldots, X_n, X_{n+1}$. For convenience, let $x_{(0)} = -\infty$ and $x_{(n+1)} = \infty$, with $x_{(0)} = 0$ for non-negative random variables. These n observations divide the real line into $n+1$ intervals: $I_j = (x_{(j-1)}, x_{(j)})$ for $j = 1, 2, \ldots, n+1$. The Hill assumption $A_{(n)}$ regarding a future observation X_{n+1}, based on these n observations, is given by:

$$P(X_{n+1} \in I_j) = \frac{1}{n+1} \quad \text{for} \quad j = 1, 2, \ldots, n+1 \tag{1}$$

NPI can be extended to predict multiple future observations by utilising a sequential version of Hill's assumption, denoted $A_{(.)}$ [17]. Let O_i represent the possible orderings of m future observations, where $i = 1, 2, \ldots, \binom{n+m}{n}$, with each ordering equally likely. For each ordering O_i, let S_j^i represent the number of future observations that fall within interval I_j, where $j = 1, 2, \ldots, n+1$. The probability of a specific ordering is:

$$P\left(\bigcap_{j=1}^{n+1} S_j^i = s_j^i\right) = P(O_i) = \frac{1}{\binom{n+m}{n}} \quad \text{for} \quad i = 1, \ldots, \binom{n+m}{n} \tag{2}$$

Here, s_j^i are non-negative integers such that $\sum_{j=1}^{n+1} s_j^i = m$. This represents the number of future observations in each interval I_j, without specifying the exact positions within each interval. The main idea of the NPI approach is to consider all possible orderings of future observations among the existing data, with each arrangement equally likely. Future observations are grouped into intervals, and we know how many will fall into each interval, though their exact values are not specified. Importantly, no further assumptions are made about the future data—each observation can take any value within its designated interval.

NPI is a powerful tool in imprecise probability theory, providing bounds for probabilities based on Hill's assumption $A_{(n)}$. Although NPI does not yield exact probabilities, it offers useful bounds, especially in cases of limited or imprecise data [5, 14]. The method calculates lower probabilities by counting all orderings where an event must occur, while upper probabilities are based on all orderings where the event is possible [5]. NPI has been successfully used in objective (Bayesian) inference [9], addressing problems that require probabilistic predictions rather than precise values. Additionally, NPI is *exactly calibrated* [26], a property ensuring that the inferences made are consistent with empirical probabilities.

In the context of reproducibility, statistical reproducibility refers to the probability of obtaining the same test outcome when a hypothesis test is repeated under identical conditions. In NPI, the process begins by performing a hypothesis test on an original sample of size n. Based on the test statistic, we determine whether to reject the null hypothesis H_0 or not. Next, we predict a future sample of size n under the assumption that all possible orderings of the n future observations among the n existing data points are equally likely. For each such ordering, we assess whether H_0 is certainly rejected, possibly rejected, or possibly not rejected. By examining these different orderings, we calculate the lower reproducibility probability by counting the number of orderings where the conclusion is certainly the same as the one derived from the original test. For the upper reproducibility probability, we extend this count to include the 'possibly' cases, where the conclusion is potentially consistent with the original test.

We should note that the sample sizes for the original and future tests do not necessarily have to be the same. However, in the reproducibility setting, assuming $n = m$ is a natural choice. This ensures that the hypothesis test is conducted under identical conditions, allowing for a fair comparison between the original test outcome and the outcomes of repeated tests on future data.

However, as sample sizes grow, the NPI-RP approach becomes computationally expensive. The number of possible future arrangements increases exponentially, so even with a relatively small sample size of 16, the number of potential orderings can quickly become prohibitively large. To address this challenge, Coolen and Marques [16] introduced a sampling methodology, where instead of calculating all possible orderings, they suggest randomly sampling future data arrangements. This method ensures that each future arrangement is equally likely and that each selection is independent of the others. By using a sufficiently large number of samples, the difference between sampling with and without replacement becomes negligible, making the approach computationally feasible. This allows lower and upper reproducibility probabilities to be estimated without the need to evaluate all possible orderings.

Another approach to improve computational efficiency is the NPI bootstrap (NPI-B) method, introduced by Coolen and Binhimd [16]. This resampling-based technique simplifies the calculation of reproducibility probabilities for various nonparametric tests. While the exact computation of reproducibility probabilities can be intensive, NPI-B provides a practical alternative by estimating values within the lower and upper bounds [12, 13]. Unlike traditional bootstrap methods, NPI-B is specifically designed for predicting future observations, offering a unique perspective within frequentist statistics.

The steps to generate an NPI-B sample for one-dimensional real-valued data in the NPI-B method are as follows [12, 13]:

1. Select an interval $I_j = (x_{(j-1)}, x_{(j)})$ where $j = 1, 2, \ldots, n+1$.
2. If I_j is finite, sample a future observation uniformly from this interval.
3. If I_j is an open-ended interval ($(-\infty, x_{(1)})$ or $(x_{(n)}, +\infty)$), sample using the tails of a normal distribution. The mean μ is set to $\frac{x_{(1)}+x_{(n)}}{2}$, and the standard deviation σ is set to $\frac{x_{(n)}-\mu}{\Phi^{-1}(\frac{n}{n+1})}$, where Φ^{-1} is the inverse of the standard normal distribution function. In the case of non-negative data $(0, +\infty)$, if the chosen interval is $(x_{(n)}, +\infty)$, sample the future value from the tail of an exponential distribution with rate $\lambda = \frac{\ln(n+1)}{x_{(n)}}$.
4. Add the newly sampled observation to the original data, creating a new dataset of size $n+1$.
5. Repeat steps 1-4 to generate a total of m future observations, forming one NPI-B sample.
6. Repeat steps 1-5 N times to create multiple NPI-B samples.

In Step 3, sampling from the tails of a fitted normal (or exponential) distribution is used to handle open-ended intervals, as proposed by Coolen and Bin Himd [13]. Within the NPI-B framework, standard normal and exponential distributions are employed to address the challenge of sampling from such unbounded intervals, where uniform sampling is not feasible. Specifically, the tail of a standard normal distribution is used for data defined on the entire real line, and the tail of an exponential distribution is used for data restricted to non-negative values. These distributional choices align with the probabilistic structure of NPI-B, particularly the assignment of a $1/(n+1)$ probability mass to the open-ended intervals.

The NPI-B framework can be applied to estimate reproducibility probability (NPI-B-RP) by iterating the statistical test multiple times and assessing the consistency of outcomes [8]. Algorithm 1 outlines the process for calculating NPI-B-RP for any test based on the methods derived from the NPI-B framework for the Wilcoxon-Mann-Whitney test (WMT) [8] and t-test [35].

The simulations were conducted using the statistical software R (version 4.2.2) on a standard laptop equipped with a 2.6 GHz Intel Core i7 processor. The average runtime for 100 repetitions of an NPI-B simulation with a sample size of 1000, used

Algorithm 1 NPI-B-RP for a statistical test

```
Require: Original samples, $N$ (number of NPI-B samples), $h$ (number of iterations).
1: Apply the test to the original data and make a decision about $H_0$. Record $T = 1$ if $H_0$
   is rejected at significance level $\alpha$, otherwise record $T = 0$.
2: Draw $N$ NPI-B samples based on the original data, applying the same test each time.
   Record $T_j = 1$ if $H_0$ is rejected, or $T_j = 0$ if $H_0$ is not rejected.
3: Compute the Reproducibility Probability (RP):
   $$RP = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}_{(T=T_j)}$$
4: Repeat steps 2-3 for $h$ iterations, obtaining $RP_1, RP_2, \ldots, RP_h$.
```

to estimate the NPI-B-RP for a single test (sample size = 10), was approximately 1.3 minutes. Simulations with larger sample sizes required proportionally longer runtimes.

3 Reproducibility of Normality Tests

Statistical procedures, particularly parametric tests, often assume that the underlying data follow a normal distribution. This assumption is crucial for drawing reliable conclusions from these tests. To assess the validity of this assumption, several statistical tests have been developed, including the Shapiro-Wilk, Anderson-Darling, and Lilliefors tests. These tests are commonly employed in research to verify whether data meet the assumption of normality before conducting further analyses. The null hypothesis for normality tests is H_0: data follow a normal distribution, against the alternative hypothesis H_1: data do not follow a normal distribution. This study focuses on three widely used normality tests: the Shapiro-Wilk test, the Anderson-Darling test, and the Lilliefors test.

3.1 Normality tests and their statistics

The Shapiro-Wilk test is widely used to assess the normality of a dataset. Its test statistic, denoted as W, is given by [33]:

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \tag{3}$$

where $X_{(i)}$ represents the ordered data points, n is the sample size, and $\bar{X}$ is the sample mean. The constants a_i are derived from the expected values of order statistics assuming a Normal distribution.

Another common test for normality is the Anderson-Darling test. The test statistic for this method, denoted as A^2, is defined as [2, 3]:

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} \{(2i-1)\log P_{(i)} + (2n+1-2i)\log(1 - P_{(i)})\} \tag{4}$$

Here, $P_{(i)} = \Phi\left(\frac{X_{(i)} - \bar{X}}{\sigma}\right)$ is the cumulative distribution function (CDF) of the standard Normal distribution evaluated at $X_{(i)}$, and σ is the sample standard deviation.

Lastly, the Lilliefors test, another widely used method for testing normality, uses the test statistic D_n [27]:

$$D_n = \max_{i=1,\ldots,n} \left| F_n(X_i) - P_{(i)} \right| \tag{5}$$

where $F_n(X_i)$ is the empirical distribution function based on the sample, and $P_{(i)}$ is the corresponding quantile from the standard Normal distribution.

3.2 NPI-B-RP of normality tests

Despite extensive research on the performance and power of normality tests, there has been limited attention to the reproducibility of these tests. In this context, reproducibility refers to a test's ability to produce consistent results under repeated sampling from the same population. The reproducibility of normality tests can be influenced by several factors, including sample size, the underlying data distribution, and the chosen significance level. Additionally, the interplay between these variables may affect the outcomes of normality tests. To investigate these effects, we conducted simulation studies to assess the reproducibility of normality tests under different sample sizes ($n = 10, 20, 50, 100$).

Under the null hypothesis of normality (H_0), data were simulated from a Normal distribution with mean and variance both equal to 1, denoted as $N(1, 1)$. For the alternative hypothesis (H_1), we considered four non-Normal distributions: the Student's t-distribution with 3 degrees of freedom ($t(3)$), the exponential distribution with rate 1 ($Exp(1)$), and the Cauchy distribution with location parameter 0 and scale parameter 1 ($Ca(0, 1)$). We also evaluated the tests across different significance levels ($\alpha = 0.01, 0.05, 0.1$). Moreover, these simulations were conducted using data generated from a normal distribution fitted to the Sepal.Length variable of the Iris dataset.

Reproducibility was estimated using the NPI-B-RP Algorithm 1. The inputs included the original sample with sample size n, $N = 1000$ (number of NPI-B samples), and $h = 100$ (number of iterations). The number of runs per simulation was set to $K = 200$. The tests were performed with a two-sided alternative hypothesis.

The simulation results for the reproducibility of normality tests when the data were drawn from $N(1, 1)$ are shown in Figure 1. A consistent pattern was observed between the p-values and RP values. Specifically, both in the rejection and non-rejection areas, the RP values increased steadily as the p-value deviated further from the threshold. When the p-value was close to the threshold, the evidence for or against H_0 was weak, leading to low RP values. As the p-value moved further from the threshold, the evidence strengthened, resulting in higher RP values.

The RP values were influenced by sample size. For small sample sizes, RP values were higher in the non-rejection area and lower in the rejection area. As the sample size increased, the reverse trend was observed, with RP values being higher in the rejection area and lower in the non-rejection area. This behavior can be attributed to the power of normality tests, which improves with larger sample sizes. The NPI-B method, which

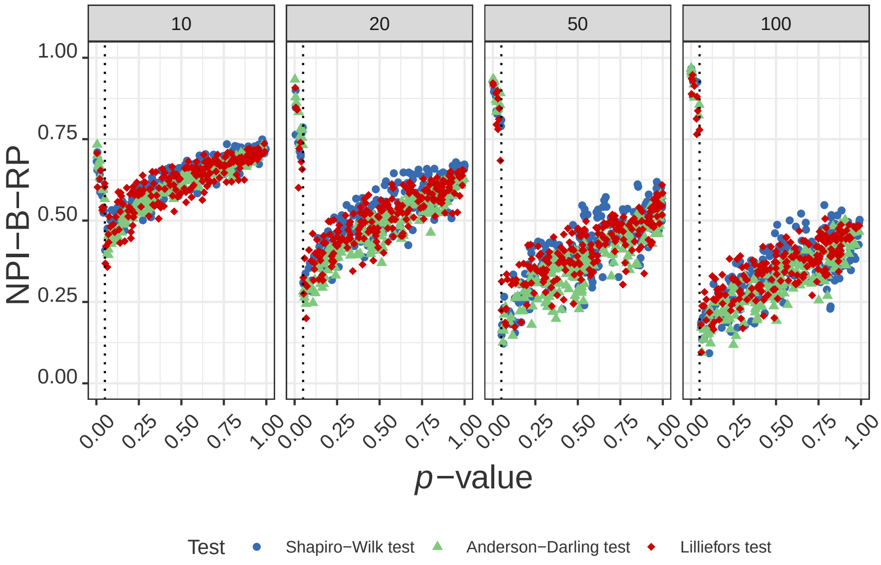

Fig. 1 Relationship between NPI-B-RP and p-value for the Shapiro-Wilk, Anderson-Darling, and Lilliefors tests when data are sampled from $N(1,1)$ at a significance level of $\alpha = 0.05$.

makes no assumptions about the distribution, results in more diverse samples, making it easier for small samples to pass normality tests. With larger sample sizes, the ability of the tests to detect deviations from normality improves, which results in lower RP values in the rejection area and higher RP values in the non-rejection area.

Interestingly, when p-values approached 1, the RP values did not approach 1. This suggests that, even when the p-value was high, there was still some uncertainty about the underlying distribution, indicating weak evidence against H_0. Consequently, the RP values did not reach close to one.

The results for the non-normal distributions also followed similar trends. When data were sampled from a $t(3)$ distribution, Figure 2 showed that the patterns observed for RP values and p-values were similar to those observed with the Normal distribution. However, a more noticeable increase in the number of samples in the rejection area was seen as the sample size increased. When data were sampled from the exponential distribution $Exp(1)$ (Figure 3), the RP values followed similar patterns to the $t(3)$ distribution, but the number of original samples in the rejection area was higher. This can be attributed to the long right tail of the exponential distribution, which increases the likelihood of extreme values and outliers. As the sample size increased, RP values in the rejection area tended to approach 1, while those in the non-rejection area decreased.

When original samples were taken from the Cauchy distribution $Ca(0,1)$, Figure 4 showed similar trends to those observed with the exponential distribution. However, the number of original samples in the rejection area and the RP values in this area were higher compared to the exponential case. This can be attributed to the Cauchy

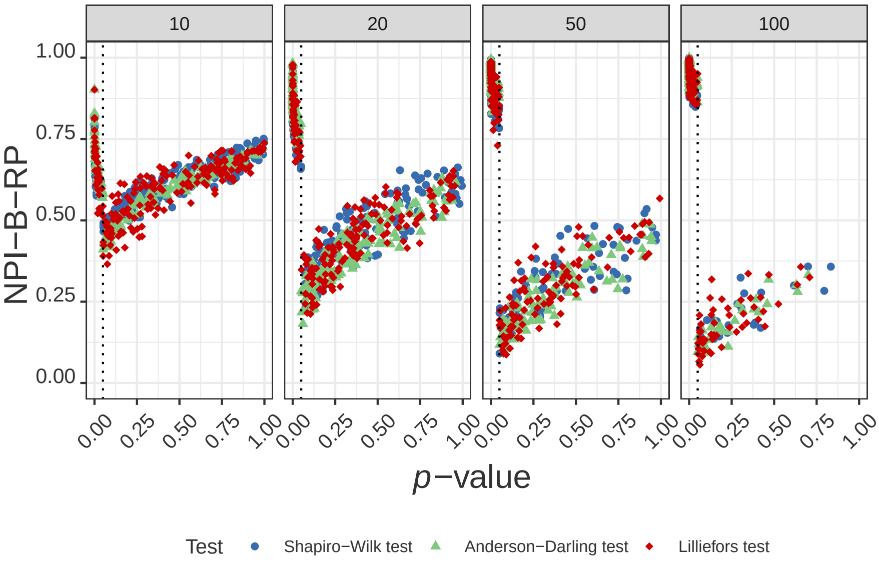

Fig. 2 Relationship between NPI-B-RP and p-value for the Shapiro-Wilk, Anderson-Darling, and Lilliefors tests when data are sampled from $t(3)$ at a significance level of $\alpha = 0.05$.

distribution's heavy tails, which make it more likely for extreme values to occur in both the original and NPI-B samples.

The effect of different significance levels ($\alpha = 0.1$, 0.05, and 0.01) on the RP values for normality tests is shown in Figure 6. As expected, RP values were generally higher at lower significance levels and decreased as the significance level increased in the non-rejection area. In the rejection area, RP values increased with higher significance levels, as higher α values made it easier to reject H_0. This occurs because a lower significance level (α) imposes a stricter threshold for rejecting H_0, making rejections less frequent. As a result, RP values tend to be lower in the rejection area and higher in the non-rejection area. Conversely, at higher significance levels, the threshold for rejecting H_0 is more lenient, leading to increased rejections, which results in higher RP values in the rejection area and lower RP values in the non-rejection area. This trend was consistently observed across all examined distributions.

Figure 5 presents the results for the reproducibility probabilities of the normality tests based on this realistic data. The patterns observed closely resemble those found in the main simulation study, reinforcing the consistency of our findings.

In summary, the simulation results demonstrate the significant impact of sample size, distribution type, and significance level on the reproducibility of normality tests. Higher sample sizes generally lead to improved power for normality tests, with RP values in the rejection area increasing and those in the non-rejection area decreasing as sample size grows. The variability of RP values also rises with increasing sample size, due to the greater complexity and diversity of NPI-B samples, which produce a wider range of possible test outcomes. Among the tests considered, the Anderson-Darling test exhibited less variability in RP values compared to the Shapiro-Wilk and Lilliefors

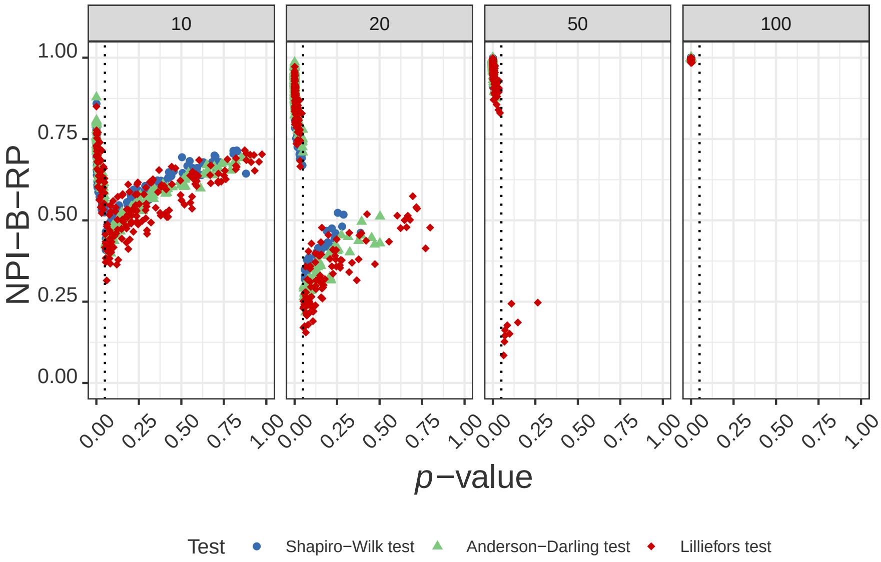

Fig. 3 Relationship between NPI-B-RP and p-value for the Shapiro-Wilk, Anderson-Darling, and Lilliefors tests when data are sampled from $Exp(1)$ at a significance level of $\alpha = 0.05$.

tests, likely due to its greater sensitivity to deviations in the tails of the distribution. Different distributions, such as the exponential and Cauchy distributions, influence the performance of normality tests, particularly in terms of the number of samples in the rejection area. The significance level also affects RP values, with lower α making the test more conservative, resulting in higher RP values in the non-rejection area and lower values in the rejection area. Conversely, higher significance levels lead to more frequent rejections of the null hypothesis, increasing RP values in the rejection area while decreasing them in the non-rejection area. This trend was consistently observed across all considered distributions. The Shapiro-Wilk test typically showed the highest RP values in the non-rejection area, while the Anderson-Darling test performed better in the rejection area.

4 Reproducibility of Equality of Variances Tests

In many statistical analyses, particularly parametric tests like ANOVA and t-tests, the assumption of equal variances across groups is essential. Ensuring homogeneity of variances enhances the accuracy of statistical inferences and interpretations. The two most commonly used tests for assessing equality of variances are the F-test and Levene's test. These tests evaluate the null hypothesis:

$$H_0 : \sigma_1^2 = \sigma_2^2 \tag{6}$$

against the alternative hypothesis:

$$H_1 : \sigma_1^2 \neq \sigma_2^2 \tag{7}$$

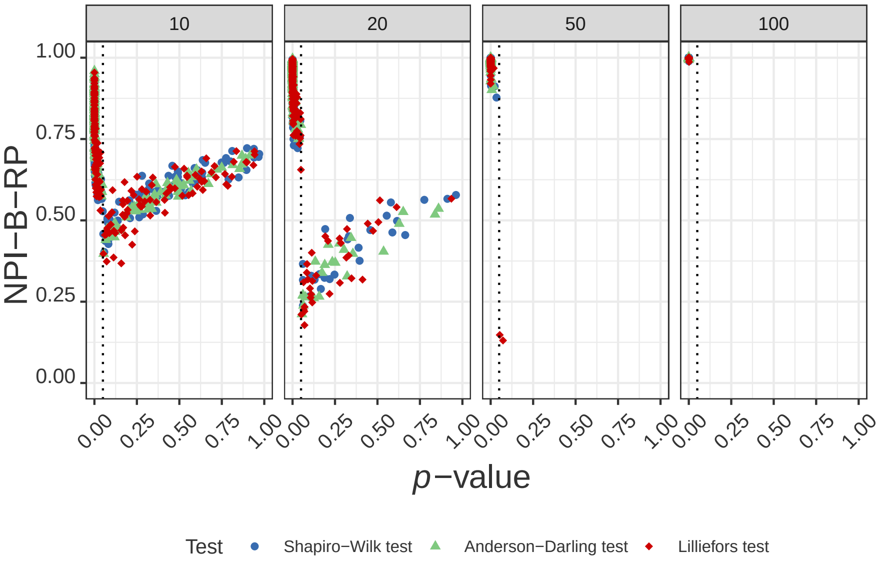

Fig. 4 Relationship between NPI-B-RP and p-value for the Shapiro-Wilk, Anderson-Darling, and Lilliefors tests when data are sampled from $Ca(0, 1)$ at a significance level of $\alpha = 0.05$.

where σ_1^2 and σ_2^2 represent the variances of the first and the second populations, respectively.

4.1 Equality of variances test statistics

The F-test evaluates the ratio of two sample variances:

$$F = \frac{S_X^2}{S_Y^2} = \frac{\sum_{i=1}^{n_X}(X_i - \bar{X})^2/(n_X - 1)}{\sum_{i=1}^{n_Y}(Y_i - \bar{Y})^2/(n_Y - 1)}, \tag{8}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means. The F-test is highly sensitive to deviations from Normality and outliers [31], which can inflate Type I error rates.

The test statistic for Levene's test is based on a one-way analysis of variance (ANOVA) using the values $Z_{ij} = |X_{ij} - \tilde{X}_i|$, where $\tilde{X}_i$ is the mean (or median) of the i-th population [25, 28]. The Levene's test statistic is given by:

$$L = \frac{(n_T - M)}{(M - 1)} \cdot \frac{\sum_{i=1}^{M} n_i(\bar{Z}_{i.} - \bar{Z}_{..})^2}{\sum_{i=1}^{M}\sum_{j=1}^{n_i}(Z_{ij} - \bar{Z}_{i.})^2} \tag{9}$$

where $\{X_{ij} : j = 1, \ldots, n_i, i = 1, \ldots, M\}$ are the samples from M populations, each with mean μ_i and variance σ_i^2 for the i-th population. n_T is the total number of observations across all groups. $\bar{Z}_{i.} = \frac{\sum_{j=1}^{n_i} Z_{ij}}{n_i}$ is the mean of the transformed data for the i-th population. $\bar{Z}_{..} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{n_i} Z_{ij}}{n_T}$ is the overall mean of the transformed data across all populations. This formulation provides the basis for evaluating the

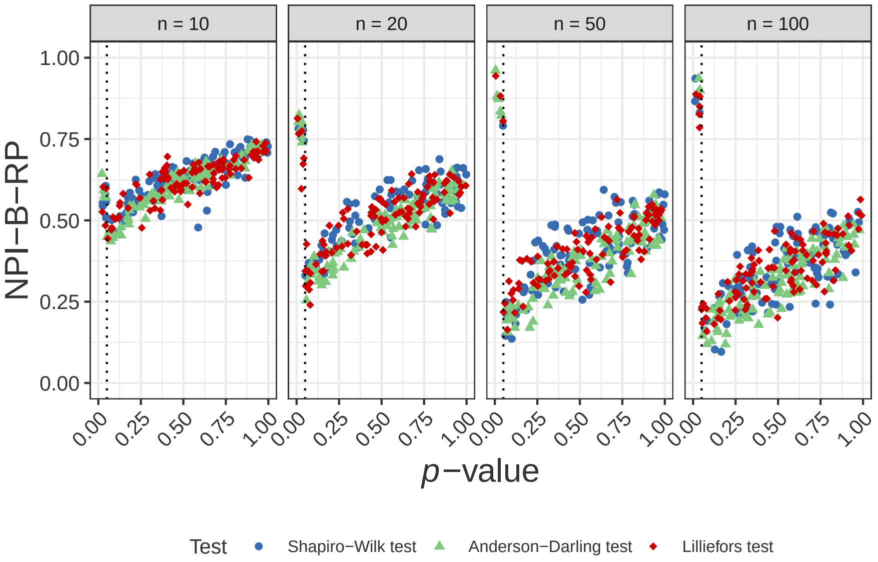

Fig. 5 Relationship between NPI-B-RP and p-value for the Shapiro-Wilk, Anderson-Darling, and Lilliefors tests using data derived from the Iris dataset at a significance level of $\alpha = 0.05$.

null hypothesis of equal variances across populations and is less sensitive to deviations from Normality compared to the F-test.

4.2 NPI-B-RP of equality of variances tests

We explore the reproducibility probability (RP) of the F-test and Levene's test by conducting simulation studies. The setup for the simulations involves generating two original samples, with $N = 1000$ and $h = 100$ for the number of replications. Each simulation run consists of $K = 200$ runs where the data is generated for sample sizes $n_1 = n_2 = 10, 25$. The tests are conducted at a 5% significance level for the two-sided equality of variances tests. We use different distributions for generating the data: Under the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$, both original samples are generated from Normal distributions with equal variances ($\sigma_1^2 = \sigma_2^2 = 1$) as $N(1, 1)$, and Exponential distributions Exp(1) with $\sigma_1^2 = \sigma_2^2 = 1$. Under the alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$, the original samples are generated from Normal distributions $N(1, 2^2)$ and $N(1, 1^2)$, where $\sigma_1^2 = 4$ and $\sigma_2^2 = 1$. Additionally, samples are generated from non-Normal distributions such as $t(3)$ and Exp(1), with variances $\sigma_1^2 = 3$ and $\sigma_2^2 = 1$. Furthermore, we also conduct simulations for the upper-tailed F-test, comparing the RP values between the two-tailed and upper-tailed F-tests.

Below are the simulation results for estimating the reproducibility probability (RP) of the F-test and Levene's test. Figure 7 displays the RP values for both tests when the original samples are drawn from a Normal distribution with equal variances ($\sigma_1^2 = \sigma_2^2 = 1$). RP values are low when the p-values from the equality of variances tests are close to the significance level ($\alpha = 0.05$). As the p-value deviates from this threshold,

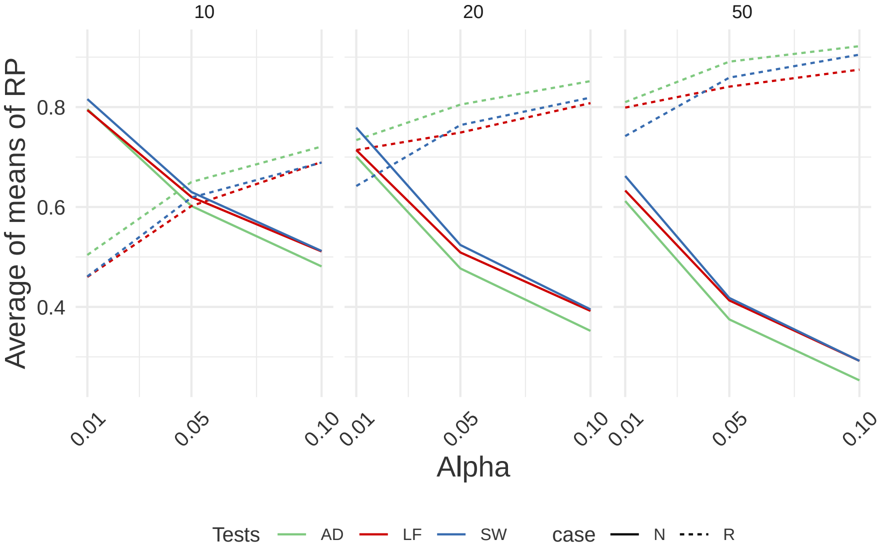

Fig. 6 Mean RP values for the Shapiro-Wilk, Anderson-Darling, and Lilliefors tests for data sampled from $N(1,1)$, with different significance levels ($\alpha = 0.01, 0.05, 0.1$) and sample sizes ($n = 10, 20, 50$). N represents the non-rejection area and R represents the rejection area.

RP values increase in both the rejection and non-rejection areas. However, even as the p-value approaches one, RP values do not fully converge to one. This is because a p-value near one suggests weak evidence against or for H_0, and does not provide certainty about the actual equality of variances. This uncertainty is reflected in the RP, which does not reach one as the p-value approaches one. RP values for Levene's test are higher than those for the F-test in the non-rejection area, but lower in the rejection area, often falling below 50%, indicating less reproducibility in this region.

Furthermore, RP values for the F-test show greater variability than those for Levene's test. This increased variability is due to the F-test's sensitivity to deviations from Normality and the inherent distributional variation in the NPI-B samples. As a result, the Type I error rate for the F-test tends to be higher than the nominal significance level (α) when Normality is violated, leading to increased variability in RP values.

Figure 8 shows RP values for the F-test and Levene's test when both original samples are drawn from a non-normal distribution with equal variances ($\sigma_1^2 = \sigma_2^2 = 1$). The RP values follow a similar pattern to those observed when both samples are from $N(1, 1^2)$, although RP values for the F-test exhibit more variability. This confirms that the F-test is highly sensitive to deviations from Normality, affecting the reproducibility probability.

Figure 9 presents RP values for the F-test and Levene's test when the original samples come from Normal distributions with different variances ($\sigma_1^2 = 4$ and $\sigma_2^2 = 1$). As the p-value approaches the threshold, RP values for both tests decrease. In the rejection area, most samples exhibit high RP values, and this trend increases as sample

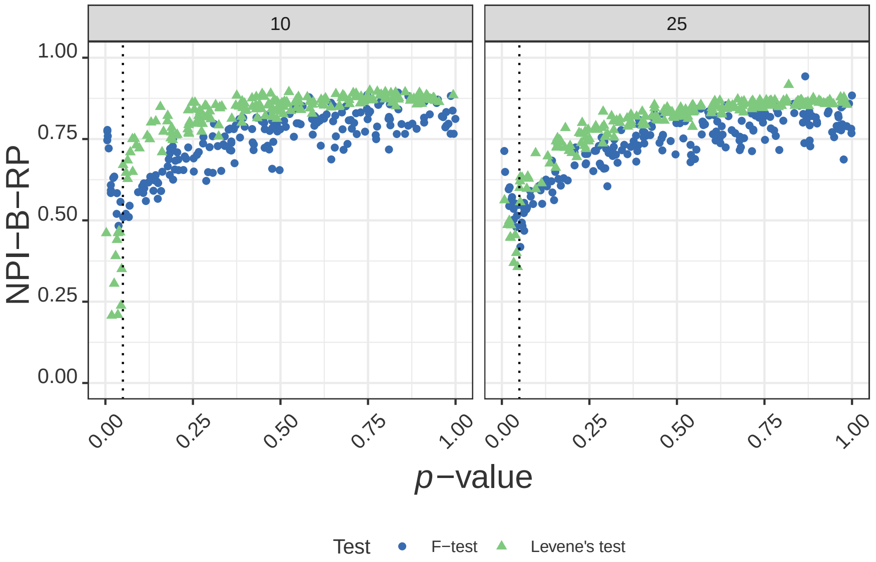

Fig. 7 Relationship between NPI-B-RP and p-value for the F-test and Levene's test under H_0, with both samples drawn from $N(1, 1^2)$.

size grows. The number of samples that apply Levene's test is smaller than that of the F-test in the non-rejection area.

Figure 10 demonstrates the RP values for the F-test and Levene's test when the original samples are drawn from a non-Normal distribution with different variances ($\sigma_1^2 = 3$ and $\sigma_2^2 = 1$). As expected, RP values increase as the p-value moves further from the threshold. Most original samples are located in the rejection area, with the number of such samples increasing as sample size grows. While there is variability in the F-test's RP values in the non-rejection area, this variability diminishes in the rejection area. This may be due to a wider range of data distributions with varying degrees of variance inequality still leading to non-rejection of H_0 in the non-rejection area, resulting in greater variability in RP values. However, once H_0 is rejected, the range of potential results becomes more constrained, reducing the variability in RP values.

Figures 11, 12, 13, and 14 compare RP values for the two-sided and upper-sided F-tests. RP values for the upper-sided F-test are closer to one when the p-value approaches one, in comparison to the two-sided tests. Additionally, RP values for the upper-sided F-test exhibit less variability, particularly when dealing with non-Normal distributions. This can be attributed to the nature of the alternative hypothesis in the upper-sided F-test, which tests whether the variance of one population is significantly higher than that of another, without considering the possibility of the reverse. This more focused hypothesis leads to a narrower range of potential results, resulting in reduced variability in test outcomes. In contrast, the two-sided test considers both

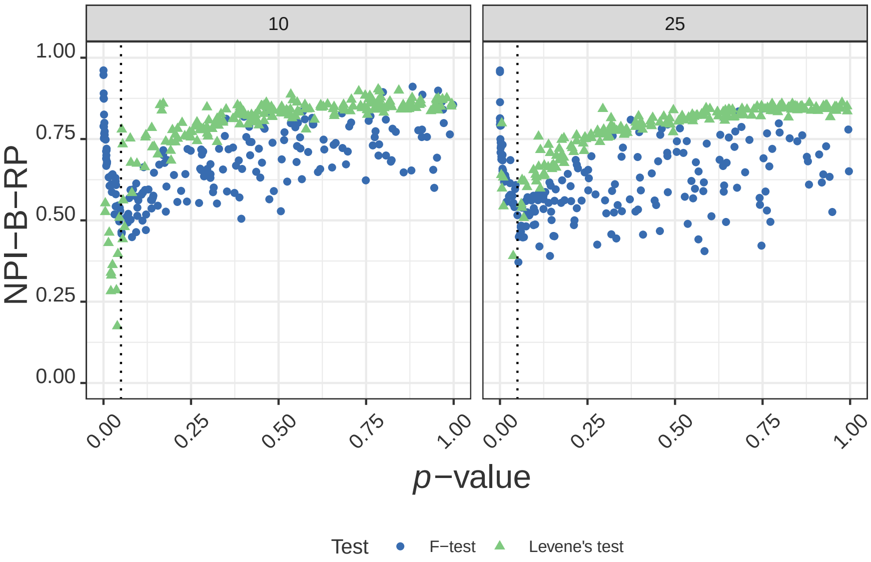

Fig. 8 Relationship between NPI-B-RP and p-value for the F-test and Levene's test under H_0, with samples drawn from $Exp(1)$.

possibilities of unequal variances, leading to a wider range of potential results and, consequently, more variability in RP values.

In conclusion, the simulation results reveal that the F-test is highly sensitive to deviations from Normality, resulting in substantial RP variability, particularly under non-Normal distributions. Levene's test shows greater RP stability in the non-rejection area but lower RP in the rejection area compared to the F-test. Increasing sample size improves RP in the rejection area but increases RP variability in the non-rejection area. Lastly, the upper-tailed F-test exhibits higher RP and lower variability, particularly in non-Normal settings, making it more reliable under such conditions.

5 Conclusions

This study examined the reproducibility probability (RP) for two types of hypothesis tests: normality tests (Shapiro-Wilk, Anderson-Darling, and Lilliefors) and tests for equality of variances (F-test and Levene's test), using the nonparametric predictive inference (NPI) bootstrap method to estimate RP.

The simulation results revealed several key insights. For normality tests, RP values decreased as p-values approached the significance threshold. Sample size played a crucial role: as the sample size increased, RP in the non-rejection area tended to decrease, while RP in the rejection area increased. Larger sample sizes also led to greater variability in RP within the non-rejection area. The significance level influenced RP as well, with higher levels corresponding to lower RP in the non-rejection area and higher RP in the rejection area. Among the normality tests, the Shapiro-Wilk

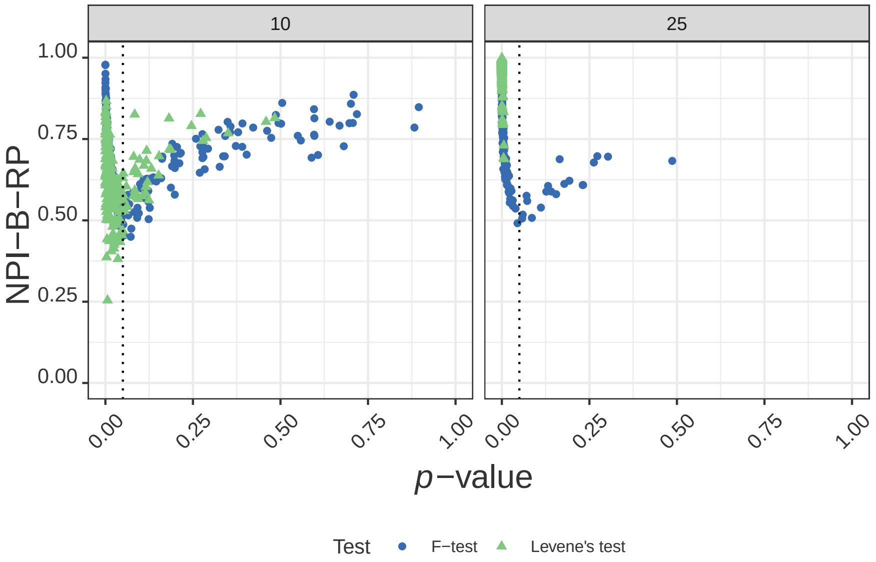

Fig. 9 Relationship between NPI-B-RP and p-value for the F-test and Levene's test under H_1, with samples drawn from $N(1, 2^2)$ and $N(1, 1^2)$.

test typically exhibited the highest RP in the non-rejection area, while the Anderson-Darling test showed the highest RP in the rejection area.

For the equality of variances tests, RP patterns resembled those observed in the normality tests. However, the F-test exhibited greater variability, especially when samples were drawn from non-normal distributions. Notably, RP values for the upper-sided F-test tended to converge towards 1 more rapidly than those for the two-sided F-test.

Future research will focus on examining the impact of pre-testing (e.g., normality and equality of variances tests) on the reproducibility of subsequent two-sample location tests, with a manuscript currently in progress. Another important area for exploration will be addressing situations with low statistical reproducibility—whether by increasing sample sizes, exploring alternative tests with better reproducibility, or refining experimental designs. Additionally, investigating the reproducibility of other pre-tests, such as additional normality tests, tests for equality of variances, independence tests, and symmetry tests, could provide valuable insights for improving the overall reliability of statistical analyses. These topics, along with others, remain open for further investigation.

Beyond its methodological contributions, the proposed NPI-B-RP approach offers practical value for applied research. By quantifying the reproducibility of hypothesis test outcomes without strong parametric assumptions, our method enables researchers to assess the stability of statistical decisions under uncertainty. This is particularly beneficial in contexts where sample sizes are small or p-values are near conventional significance thresholds. In such settings—common in disciplines like psychology, biology, and medicine—our approach can enhance the transparency and robustness of inference, ultimately supporting more reliable and reproducible scientific findings.

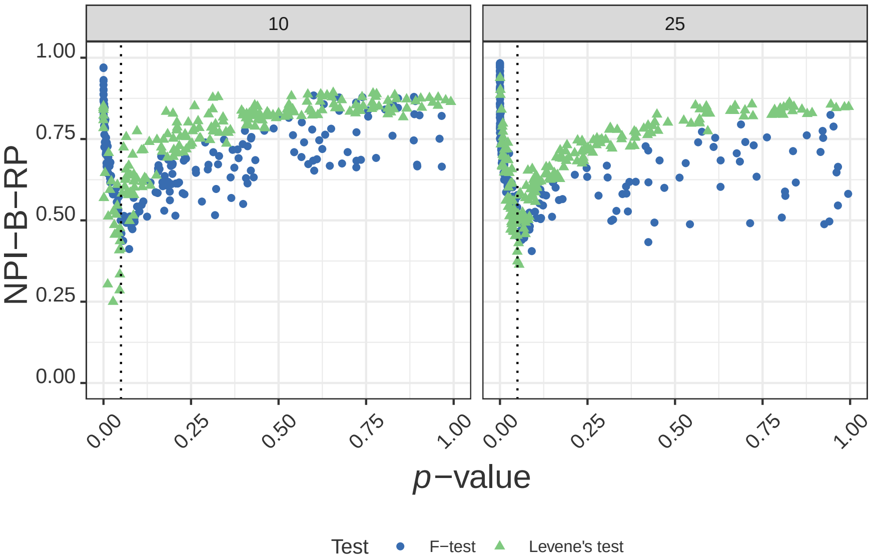

Fig. 10 Relationship between NPI-B-RP and p-value for the F-test and Levene's test under H_1, with samples drawn from $t(3)$ and $Exp(1)$.

Conflict of interest. The author states that there is no conflict of interest.

Acknowledgements. This research was conducted as part of the first author's PhD study at Durham University. The first author would like to thank the University of Bisha, Saudi Arabia, and the Saudi Cultural Bureau in London for their financial support through the PhD studentship.

References

[1] Alqifari, H. N. (2017). *Nonparametric Predictive Inference for Future Order Statistics*. PhD thesis, Durham University.

[2] Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2):193–212.

[3] Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769.

[4] Atmanspacher, H. and Maasen, S. (2016). *Reproducibility: principles, problems, practices, and prospects*. John Wiley & Sons, Hoboken, NJ.

[5] Augustin, T. and Coolen, F. P. A. (2004). Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, 124(2):251 – 272.

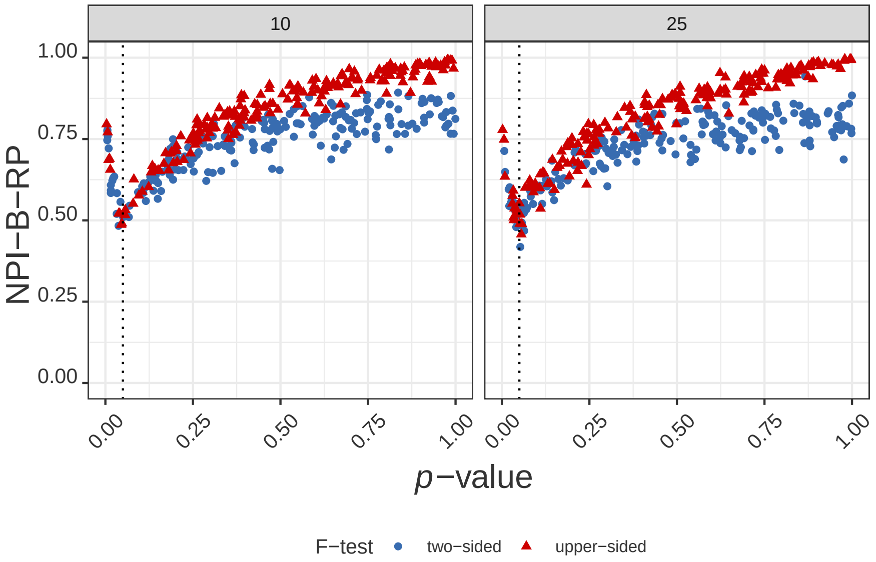

Fig. 11 Comparison of RP values between two-sided and upper-sided F-tests under H_0, with both samples drawn from $N(1, 1^2)$.

[6] Baker, R. M., Coolen-Maturi, T., and Coolen, F. P. A. (2017). Nonparametric predictive inference for stock returns. *Journal of Applied Statistics*, 44(8):1333–1349.

[7] Billheimer, D. (2019). Predictive inference and scientific reproducibility. *The American Statistician*, 73(sup1):291–295.

[8] BinHimd, S. (2014). *Nonparametric Predictive Methods for Bootstrap and Test Reproducibility*. PhD thesis, Durham University.

[9] Coolen, F. P. A. (2006). On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language, and Information*, 15(1/2):21–47.

[10] Coolen, F. P. A. and Alqifari, H. N. (2018). Nonparametric predictive inference for reproducibility of two basic tests based on order statistics. *REVSTAT: Statistical Journal*, 16(2):167–185.

[11] Coolen, F. P. A. and Alqifari, H. N. (2019). Robustness of nonparametric predictive inference for future order statistics. *The Journal of Statistical Theory and Practice*, 13:12.

[12] Coolen, F. P. A. and BinHimd, S. (2014). Nonparametric predictive inference for reproducibility of basic nonparametric tests. *Journal of Statistical Theory and Practice*, 8(4):591–618.

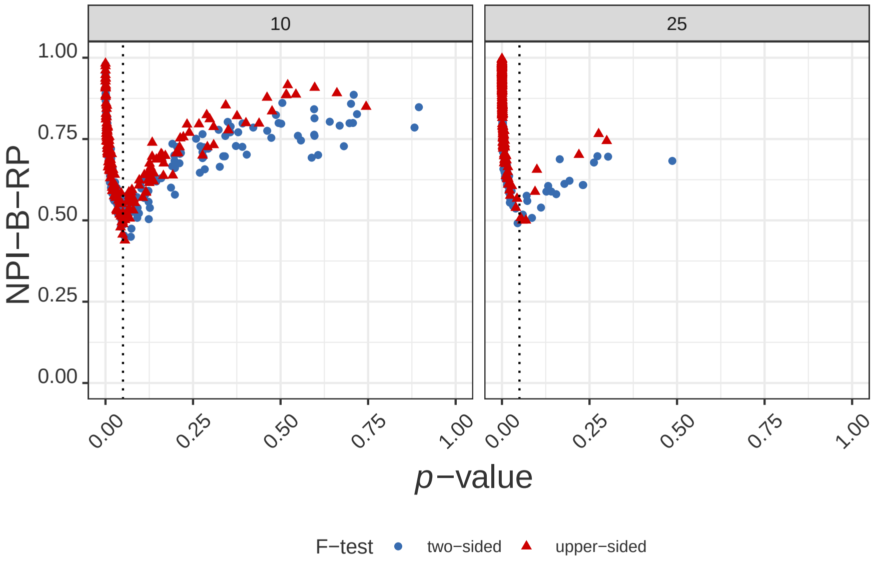

Fig. 12 Comparison of RP values between two-sided and upper-sided F-tests under H_1, with samples drawn from $N(1, 2^2)$ and $N(1, 1^2)$.

[13] Coolen, F. P. A. and BinHimd, S. (2020). Nonparametric predictive inference bootstrap with application to reproducibility of the two-sample kolmogorov–smirnov test. *Journal of Statistical Theory and Practice*, 14(2):1–13.

[14] Coolen, F. P. A. and Coolen-Maturi, T. (2025a). Nonparametric predictive inference. In Lovric, M., editor, *International Encyclopedia of Statistical Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2nd edition edition. to appear.

[15] Coolen, F. P. A. and Coolen-Maturi, T. (2025b). Statistical reproducibility. In Lovric, M., editor, *International Encyclopedia of Statistical Science*. Springer, Heidelberg, 2nd edition edition. to appear.

[16] Coolen, F. P. A. and Marques, F. J. (2020). Nonparametric predictive inference for test reproducibility by sampling future data orderings. *Journal of Statistical Theory and Practice*, 14:1–22.

[17] Coolen-Maturi, T., Coolen, F. P. A., and Alqifari, H. (2018). Non-parametric predictive inference for future order statistics. *Communications in Statistics - Theory and Methods*, 47(10):2527–2548.

[18] Coolen-Maturi, T., Elkhafifi, F. F., and Coolen, F. P. A. (2014). Three-group ROC analysis: A nonparametric predictive approach. *Computational Statistics & Data Analysis*, 78:69–81.

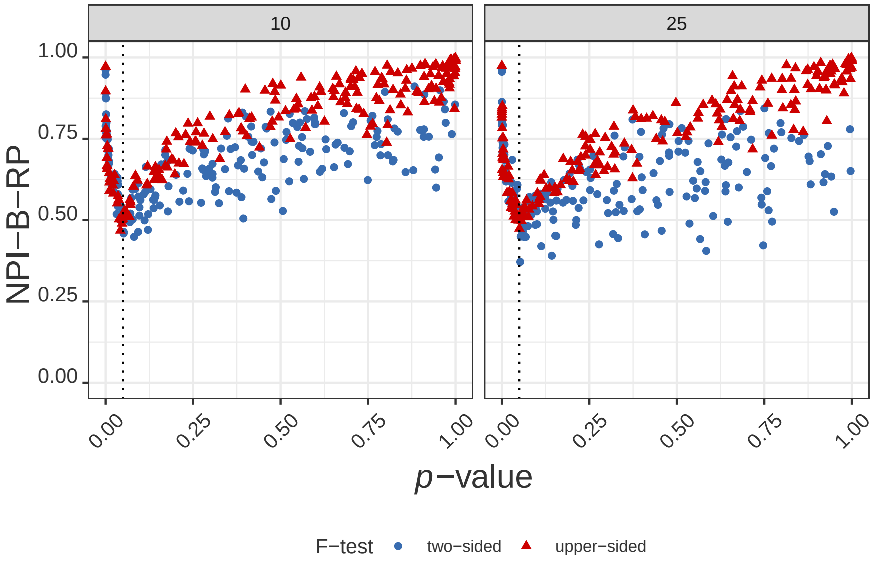

Fig. 13 Comparison of RP values between two-sided and upper-sided F-tests under H_0, with samples drawn from $Exp(1)$.

[19] De Capitani, L. and De Martini, D. (2011). On stochastic orderings of the Wilcoxon rank sum test statistic—with applications to reproducibility probability estimation testing. *Statistics & Probability Letters*, 81(8):937–946.

[20] De Capitani, L. and De Martini, D. (2015). Reproducibility probability estimation and testing for the Wilcoxon rank-sum test. *Journal of Statistical Computation and Simulation*, 85(3):468–493.

[21] De Capitani, L. and De Martini, D. (2016). Reproducibility probability estimation and RP-testing for some nonparametric tests. *Entropy*, 18(4):142.

[22] De Martini, D. (2008). Reproducibility probability estimation for testing statistical hypotheses. *Statistics & Probability Letters*, 78(9):1056–1061.

[23] Goodman, S. N. (1992). A comment on replication, p-values and evidence. *Statistics in Medicine*, 11(7):875–879.

[24] Hill, B. M. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, 63(322):677–691.

[25] Kennedy, J. J. and Bush, A. J. (1985). *An Introduction to the Design and Analysis of Experiments in Behavioral Research*. University Press of America.

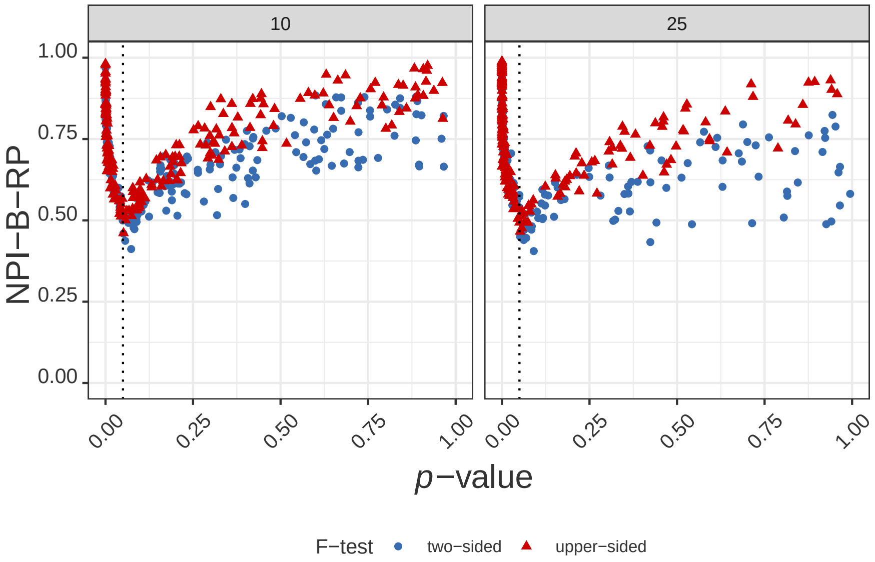

Fig. 14 Comparison of RP values between two-sided and upper-sided F-tests under H_1, with samples drawn from $t(3)$ and $Exp(1)$.

[26] Lawless, J. F. and Fredette, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika*, 92:529–542.

[27] Lilliefors, H. W. (1967). On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402.

[28] Lim, T. S. and Loh, W. Y. (1996). A comparison of tests of equality of variances. *Computational Statistics & Data Analysis*, 22(3):287–301.

[29] Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review*, 16:617–640.

[30] of Sciences, N. A., Medicine, Policy, Affairs, G., on Research Data, B., Information, on Engineering, D., Sciences, P., on Applied, C., Statistics, T., et al. (2019). *Reproducibility and replicability in science*. National Academies Press.

[31] Rasch, D. and Guiard, V. (2004). The robustness of parametric statistical methods. *Psychology Science*, 46:175–208.

[32] Senn, S. (2002). A comment on 'a comment on replication p-values and evidence'. *Statistics in Medicine*, 21(16):2437–2444.

[33] Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

[34] Simkus, A. (2023). *Contributions to Statistical Reproducibility and Small-Sample Bootstrap.* PhD thesis, Durham University.

[35] Simkus, A., Coolen, F. P. A., Coolen-Maturi, T., Karp, N. A., and Bendtsen, C. (2022). Statistical reproducibility for pairwise t-tests in pharmaceutical research. *Statistical Methods in Medical Research*, 31(4):673–688.

[36] Simkus, A., Coolen-Maturi, T., Coolen, F. P. A., and Bendtsen, C. (2025). Statistical perspectives on reproducibility: Definitions and challenges. *Journal of Statistical Theory and Practice*, 19(3):40.